U7

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : C12P 21/06, 19/34, C12N 15/00, 5/00, 1/00 | A1

(11) International Publication Number: WO 94/16092

(43) International Publication Date: 21 July 1994 (21.07.94)

(21) International Application Number: PCT/US94/00254

(22) International Filing Date: 5 January 1994 (05.01.94)

(30) Priority Data:
08/000,619 5 January 1993 (05.01.93) US

(71)(72) Applicant and Inventor: JARVIK, Jonathan, Wallace [US/US]; 6419 Beacon Street, Pittsburgh, PA 15217 (US).

(81) Designated States: European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: METHOD FOR PRODUCING TAGGED GENES, TRANSCRIPTS, AND PROTEINS

(57) Abstract

The invention described here is a method whereby a molecular tag is put on a gene, transcript and protein in a single recombinational event. The protein tag takes the form of a unique peptide that can be recognized by an antibody or other specific reagent, the transcript tag takes the form of the sequence of nucleotides encoding the peptide that can be recognized by a specific polynucleotide probe, and the gene tag takes the form of a larger sequence of nucleotides that includes the peptide-encoding sequence and other associated nucleotide sequences. The central feature of the invention in its essential form is that the tag-creating DNA has a structure such that when it is inserted into an intron within a gene it creates two hybrid introns separated by a new exon encoding the protein tag. A major utility of the method is that it allows one to identify new proteins or protein-containing structures, and, having done so, to readily identify and analyze the genes encoding those proteins.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgystan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo			SE	Sweden
CH	Switzerland	KR	Republic of Korea	SI	Slovenia
CI	Côte d'Ivoire	KZ	Kazakhstan	SK	Slovakia
CM	Cameroon	LI	Liechtenstein	SN	Senegal
CN	China	LK	Sri Lanka	TD	Chad
CS	Czechoslovakia	LU	Luxembourg	TG	Togo
CZ	Czech Republic	LV	Latvia	TJ	Tajikistan
DE	Germany	MC	Monaco	TT	Trinidad and Tobago
DK	Denmark	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	US	United States of America
FI	Finland	ML	Mali	UZ	Uzbekistan
FR	France	MN	Mongolia	VN	Viet Nam
GA	Gabon				

METHOD FOR PRODUCING TAGGED GENES, TRANSCRIPTS, AND PROTEINS.

Background - Field of Invention

This invention relates to the fields of Molecular Biology and Molecular Genetics with specific reference to the identification and isolation of proteins and of the genes and transcripts that encode them.

Background - Description of Prior Art

The primary area of the invention - the identification and tagging of genes and proteins - has received a great deal of attention, and many successful methods have been devised. None of these methods, however, has the feature of tagging gene, transcript and protein in a single event.

Linkage Analysis. Genes have traditionally been identified by identifying mutations and then mapping them with respect to one another by means of genetic crosses. This kind of mapping, or linkage analysis, does not serve to isolate the genes themselves nor does it indicate anything about the genes' molecular structure or function. In recent years a form of linkage analysis using restriction fragment length polymorphisms (RFLPs) has come into use (1). This method serves to identify DNA sequences that are linked to a gene of interest, and, having identified such a DNA sequence, it is possible in principle, and sometimes in practice, to identify and clone the gene itself

by performing chromosome walks or jumps (2). It should be stressed that. even when successful, this strategy identifies the gene. not the protein encoded by the gene.

Transposon Tagging. Another technique for cloning genes that has been developed relatively recently goes by the name transposon tagging. In this technique (3). mutations due to the insertion of transposable elements into new sites in the genome are identified. and the genes in which the transposons lie can then cloned using transposon DNA as a molecular probe. Transposon tagging, like RFLP/linkage analysis. identifies genes, not proteins.

Enhancer trapping. Another method for identifying genes, enhancer trapping (4), involves the random insertion into a eucaryotic genome of a promoter-less foreign gene (the reporter) whose expression can be detected at the cellular level. Expression of the reporter gene indicates that it has been fused to an active transcription unit or that it has inserted into the genome in proximity to cis-acting elements that promote transcription. This approach has been important in identifying genes that are expressed in a cell type-specific or developmental stage-specific manner. Enhancer trapping, like like RFLP/linkage analysis and transposon tagging. identifies genes, not proteins, and it does not directly reveal anything about the nature of the protein product of a gene.

Guest Peptides and Epitope tagging. A number of studies have been performed in which new peptides have been inserted into proteins at a variety of positions by modifying the genes encoding the proteins using recombinant DNA technology. The term 'guest peptide' has been used to describe the foreign peptides in these cases. It is clear that in many cases the presence of such peptides is relatively innocuous and does not substantially compromise protein function - especially in those cases where the peptide is on the surface of the protein rather than in its hydrophobic core.

Epitope tagging (5) is a method that utilizes antibodies against guest peptides to study protein localization at the cellular level and subcellular levels. Epitope tagging begins with a cloned gene and an antibody that recognizes a known peptide (the epitope). Using recombinant DNA technology, a sequence of nucleotides encoding the epitope is inserted into the coding region of the cloned gene, and the hybrid gene is introduced into a cell by a method such as transformation. When the hybrid gene is expressed the result is a chimeric protein containing the epitope as a guest peptide. If the epitope is exposed on the surface of the protein, it is available for recognition by the epitope-specific antibody, allowing the investigator to observe the protein within the cell using immunofluorescence or other immunolocalization techniques. Epitope tagging serves to mark proteins of already-cloned genes but does not serve to identify genes.

Isolating Genes Beginning with the Proteins they Encode. A number of procedures have been developed for isolating genes beginning with the proteins that they encode. Some, such as

expression library screening (6), involve the use of specific antibodies that react to the protein of interest. Others involve sequencing all or part of the protein and designing oligonucleotide probes that can be used to identify the gene by DNA/DNA hybridization. In all of these cases, one must have specific knowledge about a protein before it is possible to take steps to clone and characterize the gene that encodes it.

cDNA Cloning and Sequencing. A method of gene identification that has received a great deal of attention in the recent past is the cloning (and in many instances,sequencing) of so-called expressed sequence tags (ESTs) from cDNA libraries made from mRNA extracted from a given tissue or cell type (7). Information about the proteins encoded by the mRNAs can be derived from the cDNA sequences by identifying and analyzing their open reading frames. In many cases such cDNAs are not full length, however, and so information about the amino-terminal portion of the protein is lacking. And, more significantly, the method tags transcript sequences and not the proteins that the transcripts encode.

RNA splicing. RNA splicing is the natural phenomenon, characteristic of all eucaryotic cells, whereby introns are removed from primary RNA transcripts. A large body of research has revealed that an intron is functionally defined by three components - a 5' donor site, a branch site and a 3' acceptor site (8). If these sites are present, and if the intron is not too large (it can be at least as large as 2kb in many organisms), and if the distance between the branch and 3' acceptor sites is appropriate, the cellular splicing machinery is activated and the intron is removed from the transcript. Many different natural DNA sequences are known to have splice site function; consensus sites for mammalian splicing are indicated in figure 1 below. Thus not only have many active splice sites been cloned, but there is a large database that can be used to design synthetic functional splice site sequences.

Gene Trapping. Gene trapping is a method used to identify transcribed genes. Gene trapping vectors carry splice acceptor sites directly upstream of the coding sequence for a reporter protein such as β-galactosidase. When the vector inserts into an intron of an actively transcribed gene, the result is a protein fusion between an N-terminal fragment of the target gene-product and the reporter protein, the activity of which is used as an indicator that integration into an active gene has occurred (9). Gene trapping seeks to identify transcribed genes - not to tag proteins, and to inactivate genes - not to produce an active tagged gene product.

Objects and Advantages of the Invention

"CD-DNA" and "CD-Tagging". The so-called central dogma of genetics states that information flows from DNA to RNA to protein. The method of this invention tags each of the classes of macromolecule included in the central dogma. Accordingly, the method is referred to herein as "CD-tagging". Likewise, the term "CD-DNA" is used herein to refer to a DNA molecule that is inserted into the genome using the method of this invention.

Identifying and Isolating Proteins, RNAs and Genes. A method that allows one to readily identify genes by observing tagged proteins has great utility relative to the prior art. CD-tagging has just this feature. In particular, when the protein tag is an epitope that is recognized by a particular antibody, cells can be treated with a CD-DNA, or with DNA constructs containing a CD-DNA, and then subjected to immunological screens or selections to identify the epitope tag. Many different screens or selections are possible, each of which has its own particular advantages. These include direct or indirect immunofluorescence by which tagged proteins can be localized to particular regions or subcellular structures within a cell, immunoblot analysis by which the abundance, molecular weight and isoelectric points of tagged proteins can be determined, enzyme linked immuno-assays (ELISAs) by which internal or secreted tagged proteins can be distinguished, and fluorescence-activated cell sorting (FACS) by which living cells with tagged proteins at their surfaces can be obtained.

Once proteins and genes of interest have been identified, they can be efficiently purified using standard hybridization and/or affinity-purification methods that take advantage of their specific tags.

Large Target Size in the Genome. CD-tagging depends on the insertion of a CD-DNA into an intron. Since higher eucaryotic genes contain much more intron than exon sequence, the target size is large relative to other tagging methods, in which the DNA must insert into an exon. Further, since the typical gene contains numerous introns, the boundaries of which determine the sites at which amino acid insertions in the protein can be produced by CD-tagging, it is likely that for a given protein there exist multiple sites at which peptide tags produced by CD-DNA insertions would not seriously compromise protein function. Indeed, there is evidence that the sites in many proteins that represent the exon/intron boundaries are particularly likely to be on the surface of a protein - at an ideal location to accept a guest peptide and to allow recognition of that peptide by an antibody.

Hybrid Proteins Are Expressed in Backgrounds where Normal Genes Are Also Present. As discussed earlier, experience has shown that in many, and perhaps most, cases epitope fusion proteins have normal, or nearly normal, activity. But even this is not a requirement in order for CD-tagging to have utility in identifying proteins and their genes because in many applications one or more copies of the normal gene will be present in addition to the tag-containing gene (e.g., when diploid cells are tagged); here the tagged protein need not be fully functional as long as it can, for example, co-assemble at its normal location along with the protein encoded by the unaltered gene.

Obtaining Sequence Data. Once an organism or cell line expressing a protein of interest has been identified using the method of the invention, a DNA representing a portion of mRNA encoding the protein can be obtained by standard techniques such as plasmid rescue or amplifying the sequence of interest from cDNA by means of the polymerase chain reaction (PCR) using poly-dT as one primer and a DNA complimentary to the tag-encoding sequence as the other. The amplified DNA can then be sequenced by standard methods. Knowledge of the sequence can then be used to design primers for amplification from genomic DNA in order to obtain genomic sequence information.

Utility in the Analysis of Subcellular Structures. One important application for CD-tagging is to identify proteins, and the genes encoding them, that are present in particular subcellular structures. This can be done by screening CD-DNA recipients for those that express the protein tag in the structure of interest. A significant advantage of this approach is that it does not depend upon the purification of the structure of interest, or even on the prior existence of a method for such purification, as traditional methods for characterizing subcellular structures do.

In addition to identifying proteins in known structures, CD-tagging holds the promise of identifying new structures, and the proteins they contain, that have not been explicitly recognized before.

Utility in the Analysis of Cellular Responses. CD-tagging can be used to identify proteins, and the genes encoding them, whose synthesis is stimulated by a particular treatment, such as the administration of a particular hormone or growth factor to a particular cell type. This can be accomplished by comparing treated and untreated cells to identify proteins whose levels change in response to the treatment. And, using standard immunocytochemical methods, one can discriminate among such proteins to identify those that are secreted, localized to the cell surface, or present in particular subcellular compartments.

Utility in Virology. Viral infection often leads to specific changes in cellular gene expression. Using CD-tagging, cellular genes whose expression is up or down-regulated can be identified by comparing the levels of tagged proteins in infected versus uninfected cells. Likewise, if the viral genome is tagged, the expression of viral proteins during the viral life cycle can be observed.

Utility in the Analysis of Transcriptional Regulation. Much genetic regulation occurs at the level of transcription. Because CD-tagging puts a unique tag into mRNA species derived from a tagged gene, the tag can be used to investigate mRNA synthesis and stability.

Utility in the Analysis of the Human Genome. Because most cellular functions are mediated by proteins, it is of particular interest in the context of the comprehensive analysis of the human genome to identify those parts of the genome that are expressed in the form of proteins. CD-tagging provides an efficient general method to directly identify new genes on the basis of their expression as proteins and on the basis of the location of those proteins in particular cellular or extracellular structures. In addition, CD-tagging provides a method for efficient physical and/or RFLP mapping of genes, as well as a method for the isolation of genes and transcripts via their nucleic acid tags and for the efficient purification of proteins via their epitope tags.

CD-tagging has specific advantages over the prior art method for identifying and mapping genes using expressed sequence tags (ESTs). ESTs are cDNA sequences, not genomic sequences. Thus an EST probe will hybridize not only to the true gene but to any pseudogenes that are present in the genome, thereby limiting its usefulness for mapping and cloning the true gene. Likewise, an EST probe may hybridize with closely related members of a gene family, again limiting its usefulness as a probe for a unique sequence. These limitations do not apply if a gene is identified by CD-tagging, since the method provides direct access, through the CD-DNA tag, to the true gene.

Utility in Medicine. CD-tagging has broad application to the analysis and diagnosis of disease. With regard to analysis, CD-tagging makes it possible to demonstrate, through linkage analysis, that a defect with respect to a given protein represents the primary defect for a given genetic disease or cancer. The function of the protein can then be examined in detail to gain new understanding of the biology of the disease.

With regard to diagnosis, genes that are isolated using CD-tagging can provide probes to identify disease-associated restriction fragment length polymorphisms, and they can provide primers by which mutations responsible for genetic diseases could be precisely identified. Once such polymorphisms or mutations have been identified, diagnostic tests for the presence of mutant alleles in homozygous or heterozygous individuals can be developed using standard approaches.

Likewise, proteins that are isolated using the invention can be used as antigens to develop antibodies that can be used to make molecular diagnoses for a particular genetic diseases. With regard to therapy, genes or proteins that are identified using CD-tagging could be used to treat a wide variety of infectious and non-infectious diseases.

Summary of the Invention.

The invention utilizes a "CD-DNA " molecule that contains acceptor and donor sites for RNA splicing. Between the acceptor and donor sites is a sequence of nucleotides that encodes a particular peptide (or set of three peptides, one for each possible reading frame). When the CD-DNA is inserted into an existing intron, it creates a new peptide-encoding exon surrounded by two hybrid, but functional, introns. The result is that, after transcription, RNA splicing and translation, a protein is produced that contains the peptide located precisely between the amino acids encoded by the exons that surrounded the target intron. Thus, in a single recombination event at the DNA level, 1) the gene encoding the protein is tagged by the CD-DNA sequence for recognition by a DNA probe or primer, 2) the RNA transcript encoding the protein is tagged by the peptide-encoding sequence for recognition by a DNA probe or primer, and 3) the protein is tagged by the peptide for recognition by a specific antibody or other reagent.

Figures

Figures 2-8 show the structures of a number of DNA molecules that embody the invention.The dark lines represent DNA molecules, with the thicker areas representing coding sequence. Sites in the DNA are represented by short vertical lines. Segments of each DNA are indicated below each molecule. When the DNAs are functioning when inserted into introns, transcription is from left to right for those regions where the sites are shown above the DNA molecules, and from right to left for those regions where the sites are shown below the DNA molecules.

In the figures the various DNA segments ("peptide-encoding segment", "left arm", "right arm", "central segment") are not given specific lengths. This reflects the fact that their lengths can vary considerably and need not have the same values from embodiment to embodiment. The peptide encoding segments will generally be between 24 and 75 nucleotides in length so as to encode peptides of 8 to 25 amino acids; the other segments will generally total between 100 and 1000 base pairs so that the hybrid introns created by insertion of the CD-DNA are not too large for efficient splicing. Likewise the base compositions of the various DNA segments are not defined, except at the indicated splice acceptor, branch and donor sites. These segments could be random sequences or be natural sequences without unusual structural features.

It should be emphasized that a great many different molecules of the structures claimed here can be constructed, and that a great many specific means for constructing such molecules using standard recombinant DNA technology will be obvious to an individual skilled in the arts of molecular biology.

Description of Invention.

This invention provides a method for tagging proteins and the genes and transcripts that encode them in a single recombinational event. The method involves the insertion by *in vitro* or *in vivo* recombination of a specially chosen and/or designed DNA sequence into an intron that is expressed within the genome of a cell or organism. This DNA sequence carries: 1) coding information for one or more specific peptides, typically, but not necessarily, from eight to twenty five amino acids in length, and 2) appropriately placed branch, acceptor and donor sites for RNA splicing. The nucleotide sequences representing the branch, acceptor and donor sites may represent natural sites taken from known genes or they may be rationally designed based on current knowledge of the nucleotide compositions of such sites (8).

Figures 2 through 10 show the structures of a number of different embodiments of the invention. A key and essential feature of these embodiments is that, when inserted into existing introns, they instruct the splicing machinery of the cell to recognize more than one intron where there was previously one, with these new introns flanking a new exon, or exons, encoding a peptide, or peptides, of determined amino acid sequence.

All of these embodiments can be readily produced by an individual skilled in the arts of molecular biology. I have not specified the specific means by which the embodiments are constructed because there are numerous ways, well known to an individual skilled in the arts of molecular biology, by which this can be accomplished. Likewise, I have not specified the particular nucleotide sequences present in each segment, except as specifically indicated in the text. Again, there are many sequences that could serve and that could be used by one skilled in the arts of molecular biology.

Figure 2 represents a basic embodiment of the invention. The DNA is designed to function when inserted into an intron that is transcribed from left to right. It has a peptide-encoding segment between splice acceptor donor sites. Within the left arm is a splice branch site. The size and nucleotide sequence of the peptide-encoding region determines the size and amino acid sequence of the encoded peptide, with the amino acid sequence of the peptide determined by the rules of the genetic code. The number of nucleotide pairs in the peptide-encoding region must be an even

multiple of three to ensure that the reading frame is maintained with respect to the surrounding exons.

Figure 3 shows a specific CD-DNA of the structure represented in Figure 2. and Figure 4 shows a restriction map of that same DNA molecule. The 54 bp coding sequence of this DNA encodes three distinct peptides, one of which contains an epitope recognized by a specific monoclonal antibody .

Figures 5, 6 and 7 represent embodiments designed to function when inserted into an intron in either orientation.

Figure 8 represents a circular embodiment of the invention. This embodiment could, for example. be a plasmid that contains DNA encoding the guest peptide.

Figure 9 represents an embodiment incorporating a gene, or genes, that could allow for selection in a target cell. The gene is intron-less so that it does not contribute splice sites.

Figure 10 represents a circular embodiment of the invention containing two peptide-encoding segments.

Figures 2 through 8 represent some, but by no means all, possible embodiments of the invention. More complex embodiments that retain the essential elements of the invention are also possible. For example, CD-DNAs containing more than two segments encoding guest peptides can be designed; such CD-DNAs can be relatively large and yet not lead to the generation, in the target gene, of new introns that are excessively large for efficient splicing.

It may be, for example, that in certain cells specific splice branch sites are less critical to splicing function than specific acceptor and donor sites, in which case an effective embodiment of the invention might be created without specific branch sites Furthermore, as knowledge of the biochemistry of RNA splicing accumulates in the art, DNA sequence features in addition to the known donor, branch and acceptor sites may be identified that can be included in the CD-DNA to improve the efficiency of splicing. The scope of this invention is intended to include such features.

Operation of the Invention

The design of the CD-DNA is such that when it is inserted into an existing intron, it creates, within the intron, a new peptide-encoding exon. The result is that, after transcription, RNA splicing and translation, a protein is produced that contains the peptide located precisely between

the amino acids encoded by the exons that surrounded the target intron. Thus, in a single recombination event: 1) the gene encoding the protein is tagged by the CD-DNA sequence for recognition by a DNA probe or primer, 2) the RNA transcript encoding the protein is tagged by the peptide-encoding sequence for recognition by a DNA probe or primer, and 3) the protein is tagged by the peptide for recognition by a specific antibody or other reagent.

Recombination of a CD-DNA within an intron is essential to successful CD-tagging. Figures 9 illustrates the structure of the DNA that results from the integration of a linear CD-DNA within an intron by recombination at its ends. When transcribed, this DNA yields an RNA that is spliced to produce an mRNA encoding a protein that contains a guest peptide located precisely between the protein segments encoded by the exons that bound the target intron. Figure 10 illustrates the structure of the DNA that results from the integration of a circular CD-DNA within an intron by a single crossover. When transcribed, this integrated DNA yields an RNA that is spliced to produce an mRNA encoding a protein that also contains a guest peptide (in this case encoded in two guest exons) located precisely between the protein segments encoded by the exons that bound the target intron.

Integration of a CD-DNA can be accomplished in a number of ways. One approach involves the introduction of CD-DNA into cells by standard methods such as transformation, electroporation, transfection, bulk loading, or liposome fusion, followed by nonhomologous recombination of the CD-DNA into the genome. The occurance of such recombination is well known in many cell types; sometimes the integration of foreign DNA is accompanied by a small deletion of the target sequence, but, as long as such a deletion remains within the intron, it will present no problem.

In another approach, the CD-DNA is inserted <u>in vitro</u> into a cloned DNA or a DNA library in a viral or plasmid vector by standard recombinant DNA methods, and the recombinant DNA molecules are then introduced into eucaryotic cells where the recombinant genes are expressed.

In another approach takes advantage of the mobility of transposons or retroviruses; here the CD-DNA is located on an engineered transposon or retrovirus that moves it to new sites using the recombination/transposition machinery of the transposon or retrovirus.

All of the above approaches can be (but need not necessarily be) employed without prior knowledge of the locations or structures of the target genes; in such cases, the insertion of the CD-DNA can be essentially random with respect to the recipient DNA. Integration into an intron of an expressed gene is then recognized using an appropriate detection method or methods. For example, if the guest peptide includes an epitope recognized by a specific antibody, immunofluorescence microscopy can be used to detect the presence of the tagged protein.

Insertion of the CD-DNA into recipient DNA in such a way as to allow expression in eucaryotic cells can be readily achieved by an individual skilled in the arts of molecular and cell biology using a variety of standard methods including, but not limited to, those specifically mentioned above. Many such methods are well known to an individuals skilled in the arts of molecular and cell biology.

Peptides and Epitopes

In one major class of application of CD-tagging, the peptide that is introduced into a protein is an epitope that is recognized by a specific monoclonal or polyclonal antibody. In principle, almost any amino acid sequence not present in the cells of interest could serve as such an epitope. And, while there may not be a single "optimal" epitope, epitope design could still follow a rational basis. In most cases, it would be valuable for the epitope to be on the surface of the protein where 1) it would be readily available to the antibody combining site, and 2) it would minimally disrupt the tertiary structure of the protein as a whole. Surface location can be promoted by use of hydrophilic epitopes (except in the case of integral membrane proteins, where hydrophobic epitopes can be employed). If a single repeating nucleotide is used to encode the epitope, it will yield the same poly-amino acid epitope in all three reading frames; a repeating dinucleotide will encode two potential poly-amino acid epitopes, and a repeating trinucleotide, three such epitopes. A somewhat more complex repeating sequence can be used to encode repeating di-amino acid epitopes, and still more informationally complex sequences can be used to create epitopes of a very wide variety of amino acid sequences, with the only obvious requirement being the absence of stop codons in the reading frames. Furthermore, some CD-DNAs (Figures 3,4,5) contain peptide-encoding sequences that can be read in both directions; in these cases as many as six distinct epitopes can be encoded on the same CD-DNA. Which epitope appears in the protein will then depend on the orientation the CD-DNA as well as the the reading frame that is is dictated by the specifics of the intron/exon boundaries of the target intron.

In addition to using epitopes that are designed according to the principles outlined above, other epitopes exist, such as hemaglutinin sequences from influenza virus, micro-exon 1 encoded sequence from the ubx gene of Drosophila, or sequences encoded by the myc oncogene, that have already proved their worth in epitope tagging. These very sequences can be used in embodiments of CD-tagging, thereby ensuring that the guest peptides can be identified by standard procedures.

Recipient Cells

Because RNA splicing is a universal characteristic of eucaryotic cells, CD-tagging is applicable to a very wide variety of cells and organisms, including yeasts, protozoans, algae, metazoans (both plant and animal), and somatic and germline cells derived from metazoan organisms. Because the nucleotide sequences that are necessary and sufficient for splicing are highly conserved across the eucaryotes, it is likely that in many cases the same CD-DNA will function in a variety of cell types and organisms. This is not to say, however, that a given CD-DNA will not function optimally in a given cell type or organism, and so it may prove useful to develop different CD-DNAs for use in different backgrounds. It is also the case that the signals for alternative splicing may vary from cell to cell; the optimal CD-DNA would typically be one in which splicing of the hybrid transcript always occurs. One way to maximize the liklihood of this is construct the CD-DNA using nucleotide sequences that are known to function in the very background in which the tagging is to be performed.

Identification of genes and proteins.

Once cells or organisms have been constructed by insertion of the CD-DNA sequence into the genome, or by insertion into a library that is then transferred to the genome, they can be screened with epitope-specific antibodies by standard immunological techniques. These techniques include ELISAs and western blots to identify cells in which hybrid proteins are synthesized, and immunofluorescence, immunoelectron microscopy and other immunocytochemical methods to identify the cellular and sub-cellular locations of the epitope-containing hybrid proteins. When cells or organisms carrying immunoreactive proteins in structures of interest have been identified, the epitope tags in the proteins can be used to purify them by standard affinity-based methods, and/or the epitope-encoding polynucleotide tags can be used to identify and/or sequence cDNAs made from cellular mRNA by standard methods, and/or the vector tags in the DNAs can be used to clone the genes by standard methods and/or to obtain DNA sequence data.

Experimental Results.

To test the CD-tagging method, the DNA described in Figure 3 was inserted into an intron in a well characterized gene and the gene transformed into eucaryotic cells to test for its function. The organism chosen for this purpose was the unicellular eucaryote *Chlamydomonas reinhardtii*, and the gene chosen was *pf14*, which encodes the flagellar protein RSP3. Using the genomic plasmid clone KE-RS3 (10) as a recipient, the CD-DNA of Figure 2 was inserted into a unique *Ava*III site located in intron 2 using standard recombinant DNA methods. Based on the known RSP3

sequence, it was expected that the CD-DNA would result in the insertion of a guest peptide that contained the 9-amino acid epitope recognized by monoclonal antibody 12CA5 (11). The CD-tagged plasmid DNA was then transformed into *Arg7⁻, pf14⁺* cells, and Arg⁺ transformants were examined using the polymerase chain reaction (PCR) to identify those that contained the CD-tagged DNA. Protein was prepared from fourteen such transformants, as well as from control cells that had not received the CD-DNA, and subjected to SDS slab gel electrophoresis through through 10% polyacrylamide. Included in the gels was a small epitope-tagged yeast protein as a positive control for recognition by the antibody. When western blots were prepared and probed with antibody 12CA5, a distinct signal at the expected position of RSP3 (approximately 85 kilodaltons molecular weight) was present in the lanes derived from the CD-tagged cells; no such signal was observed in the control (no CD-tagged DNA) lanes. It is concluded that the CD-DNA functioned as expected, resulting in the presence of a guest epitope in the RSP3 protein. A western blot illustrating this result is presented in Figure 12.

Conclusion, Ramifications and Scope of Invention.

In conclusion, this invention describes a method for tagging gene, transcript and protein in a single recombinational event. This method has unique and highly advantageous utility over all other methods in the prior art with similar aims.

The specific description of my invention presented above should not be construed as limitating its scope, but rather as exemplification of certain embodiments thereof. Many other variations and applications are possible. For example, peptides could be designed that have sites that lead to specific covalent modification of the tagged protein - either by a small molecule or a macromolecule. Or the peptide tag could contain a site for hydrolysis of a peptide bond by an inducible protease, thereby making it possible to assess the function of the tagged gene *in vivo*. Or CD-DNAs could contain cis-acting sites for the inducible activation of transcription arranged so that inhibitory anti-sense transcripts from the target gene are produced, thereby making it possible to assess the function of the tagged gene *in vivo*.. Or the peptide-encoding sequence could contain nucleotides that are hypermutable in vivo so as to promote mutations such as frameshifts that could inactivate protein function. Or an enhancer of transcription could be included within the CD-DNA so that expression of the target gene is stimulated by the CD-DNA. Accordingly, the scope of the invention should be determined not by the embodiments illustrated here but by the appended claims and their legal equivalents.

I claim:

1. A Method for tagging genes, transcripts and proteins comprising:

   (a) producing a DNA molecule containing within it an acceptor site for RNA splicing, and associated nucleotide sequences necessary for splice acceptor function, followed by a sequence of nucleotides encoding one or more defined peptides followed by a donor site for RNA splicing;

   (2) introducing said DNA molecule into an intron within a gene, and;

   (3) promoting expression of said gene in a cell or cells with the result that the protein product of said gene contains as part of its primary sequence a new sequence or sequences of amino acids encoded by said sequence of nucleotides.

2. The method of claim 1 wherein said DNA molecule is introduced into said intron by in vitro recombination methods.

3. The method of claim 1 wherein said DNA molecule is introduced into said intron by in vivo recombination.

4. The method of claim 1 wherein said cell is that of a procaryotic microorganism.

5. The method of claim 1 wherein said cell is that of a eucaryotic microorganism.

6. The method of claim 1 wherein said cell is a somatic cell in a multicellular organism.

7. The method of claim 1 wherein said cell is a germline cell in a multicellular organism.

8. The method of claim 1 wherein said cell belongs to a culture of pleuripotent stem cells derived from a multicellular organism.

9. The method of claim 1 wherein said cell belongs to a somatic cell culture derived from a multicellular organism.

10. The method of claim 1 wherein expression of said gene is promoted by introducing said DNA molecule into said cell by a method chosen from the following group: transformation, electroporation, transfection, bulk loading, liposome fusion.

11. The method of claim 1 wherein said DNA molecule is introduced into said intron by the method of transposon insertion.

12. The method of claim 1 wherein said DNA molecule is part of a recombinant plasmid.

13. The method of claim 1 wherein said DNA molecule is part of a recombinant virus.

14. The method of claim 1 wherein said DNA molecule is part of a recombinant transposon.

15. The method of claim 1 wherein said DNA molecule becomes stably incorporated into the genome of said cell.

16. The method of claim 1 wherein said peptide or peptides are recognized by specific monoclonal antibodies.

17. The method of claim 1 wherein said peptide or peptides are recognized by specific polyclonal antibodies.

18. The method of claim 1 wherein said peptide or peptides are recognized by specific reagents that are not antibodies.

19. The method of claim 1 wherein said DNA contains two segments, each with a structure like that described in claim 1, oriented in opposite directions.

20. The method of Claim 19 wherein said two segments share the same peptide-encoding region which can be translated in either direction to give distinct peptides.

21. The method of claim 1 wherein said DNA contains a known gene or genes located in a region outside that bounded by said branch site and 5' donor splice site.

22. The method of claim 1wherein said DNA contains multiple segments, each with a structure like that described in claim 1, oriented in the same directions.

23. The method of claim 1 wherein the DNA sequences defining the splice acceptor and donor sites are derived from the ends of natural introns.

24. Any macromolecule that is tagged using the method of this invention.

25. Any macromolecule that is discovered using the method of this invention.

26. Any subcellular or extracellular structure that is tagged using the method of this invention.

27. Any subcellular or extracellular structure that is discovered using the method of this invention.

28. Any diagnostic test for a disease employing macromolecules that are tagged using the method of this invention.

29. Any diagnostic test for a disease employing macromolecules that are discovered using the method of this invention.

30. Any therapeutic treatment for a disease employing macromolecules that are tagged or discovered using the method of this invention.

31. Any automated or semi-automated instrument that creates macromolecules tagged by the method of this invention.

32. Any automated or semi-automated instrument that identifies macromolecules tagged by the method of this invention.

33. Any cell or organism containing genes tagged by the method of this invention.

34. Any unique DNA molecule that serves to tag genes, transcripts and proteins by the method of this invention.

35. A DNA molecule of the sequence:

ATGCATGTCGACCCGGGATCCGAATTCTCGAGGCTAAGCCAGTTTTCGTAC

CCTTGACTGCGTTTCATCGATTCGCTACTAACATTGCCTTTTCCTCCTTCCCT

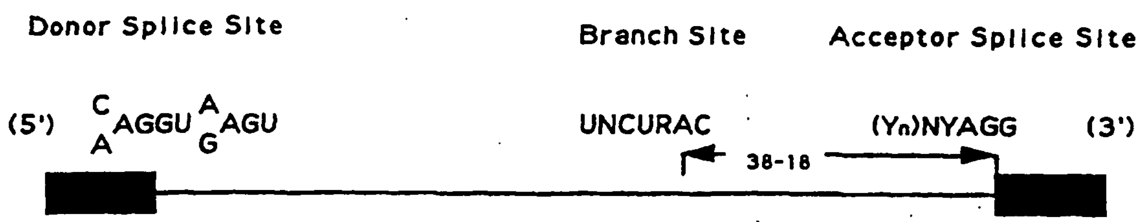
CCCACAGGTGGAAGAGCTCGGTACCCCTACGACGTCCCCGACTACGCCACGA

AGATCTCAGGTGAGTTCGCATGTGCTTCGAACTTGTGTGCATGCGTTCT

AAAAGGGCTTCTCTTGGTGTTCGATCTGGGCTAAGCTTAATTAAGAATTC

GGATCCCGGGCGTCGACATGCAT

Figure 1. Consensus sequences for splicing mammalian pre-mRNA transcripts.



R: purine; Y: pyrimidine; N: any base.

Figure 2: Structure of CD-DNA. Embodiment 1.

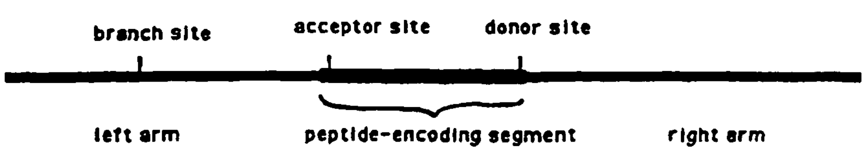


Figure 3: A specific CD-DNA.

*ATGCATGTCGACCCGGGATCCGAATTCTCGAGGCTAAGC*CAGTTTTCGTACCC

branch

TTGACTGCGTTTCATCGATTCGCTACTAACATTGCCTTTTCCTCCTTCCCT

acceptor

CCCACAG<u>GTGGAAGAGCTCGGTACCCCTACGACGTCCCCGACTACGCCACGA</u>

<u>AGATCTCAG</u>GTGAGTTCGCATGTGCTTCGAACTTGTGTGCATGCGTTCT

AAAAGGGCTTCTCTTGGTGTTCGATCTGG*GCTAAGCTTAATTAAGAATTCG*

GATCCCGGGCGTCGACATGCAT

- Polylinker (italics): 39 nucleotides.
- Right end of natural intron (derived from intron 2 of *Chlamydomonas reinhardtii* α_1 tubulin): 37 nucleotides.
- Consensus branch and acceptor sites: 36 nucleotides.
- Coding sequence (underlined): 54 nucleotides.
- Consensus donor site: 9 nucleotides.
- Left end of natural intron (derived from intron 2 of *Chlamydomonas reinhardtii* α_2 tubulin): 63 nucleotides.
- Polylinker (italics): 44 nucleotides.
- Peptides encoded by the coding sequence:

Reading frame 1:	-val-glu-glu-leu-gly-thr-pro-thr-thr-ser-pro-thr-thr-pro-arg-arg-ser-gln-
Reading frame 2:	-trp-lys-ser-ser-val-pro-leu-arg-arg-pro-arg-leu-arg-his-glu-asp-leu-
Reading frame 3:	-gly-arg-ala-arg-tyr-pro-tyr-asp-val-pro-asp-tyr-ala-thr-lys-ile-ser-

Figure 4: Physical map of the CD-DNA shown in Figure 3.

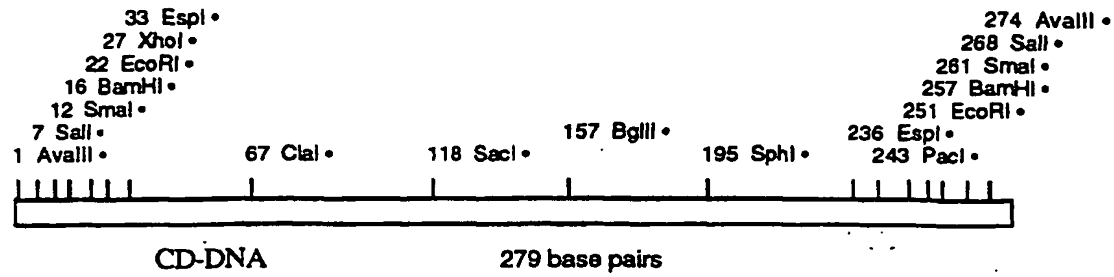


Figure 5: Structure of CD-DNA, Embodiment 2.

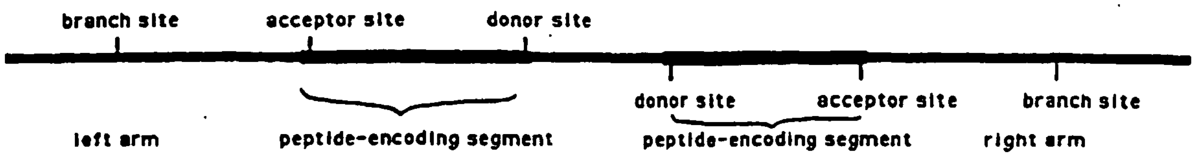


Figure 6: Structure of CD-DNA, Embodiment 3.

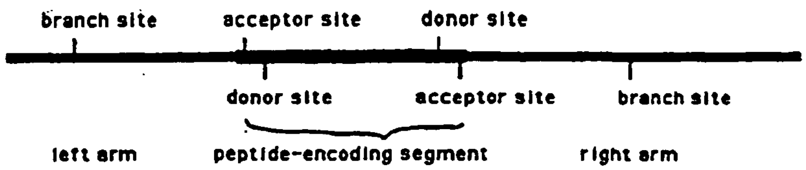


Figure 7: Structure of CD-DNA, Embodiment 4.

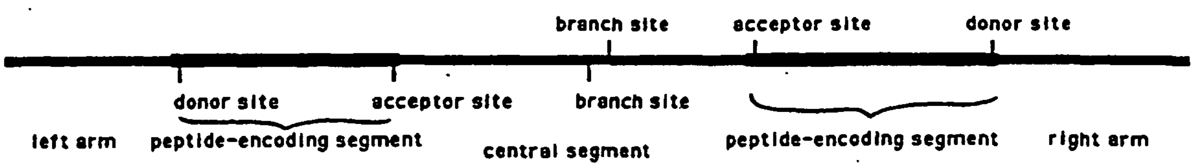


Figure 8: Structure of CD-DNA, Embodiment 5.

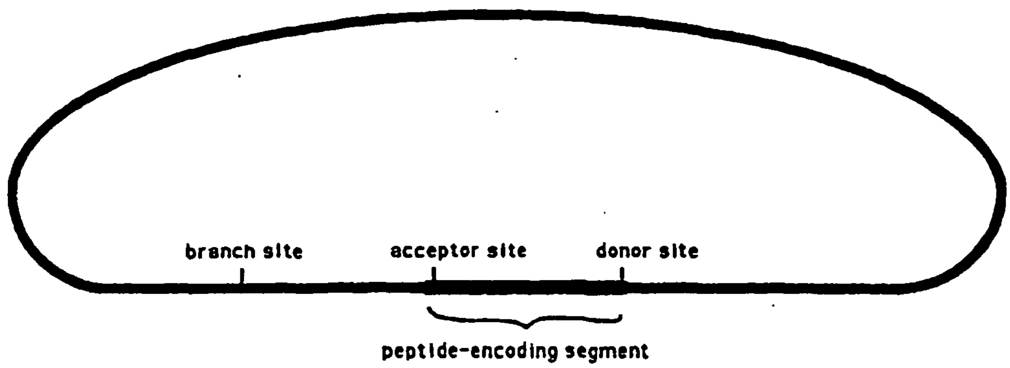


Figure 9: Structure of CD-DNA, Embodiment 6.

Figure 10: Structure of CD-DNA. Embodiment 7.

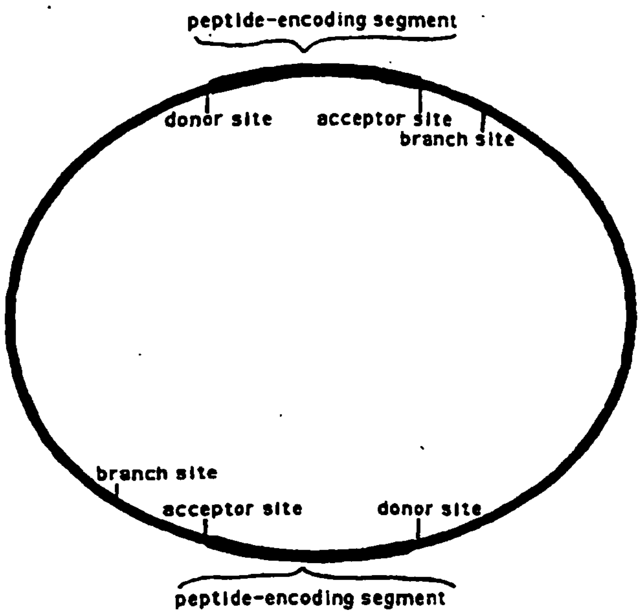
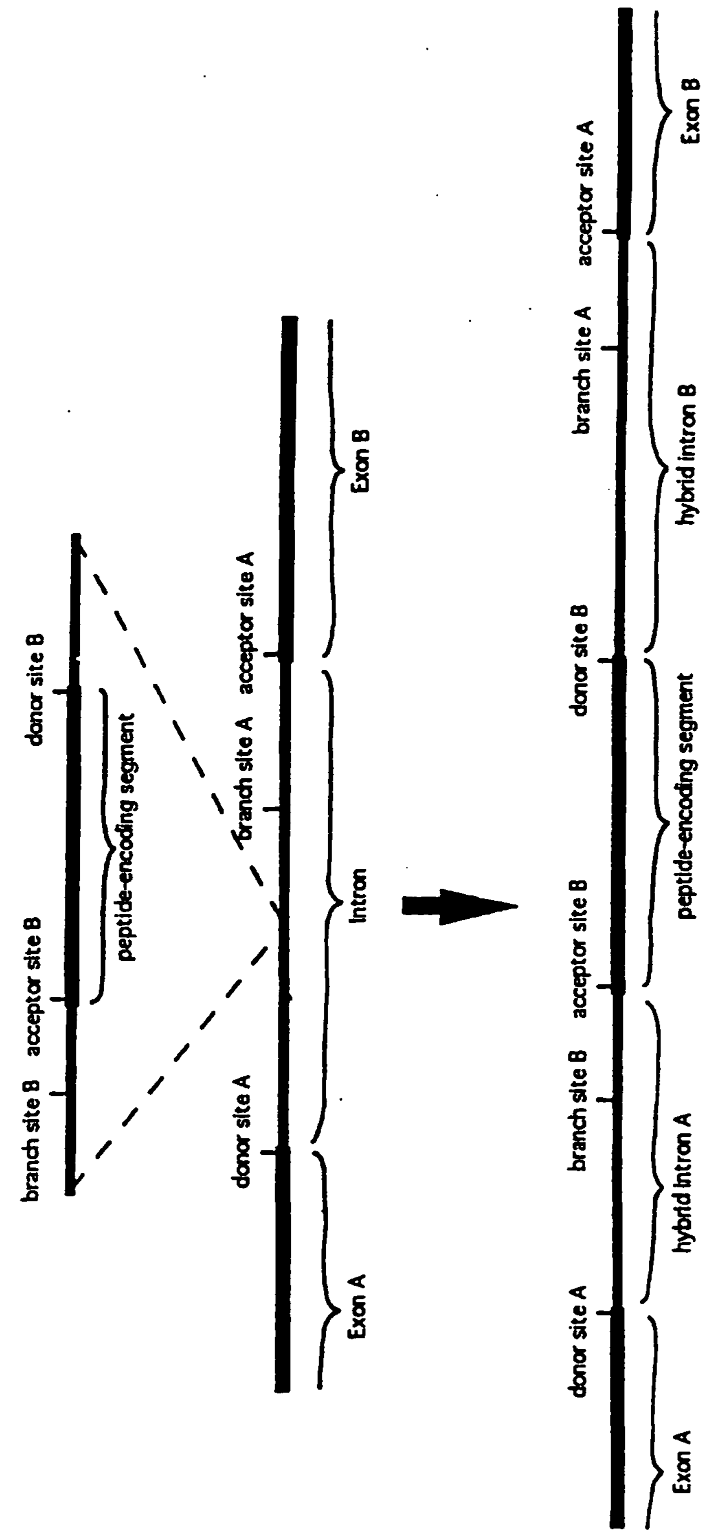


Figure 11: Structure of CD-DNA (embodiment 1) inserted into intron.



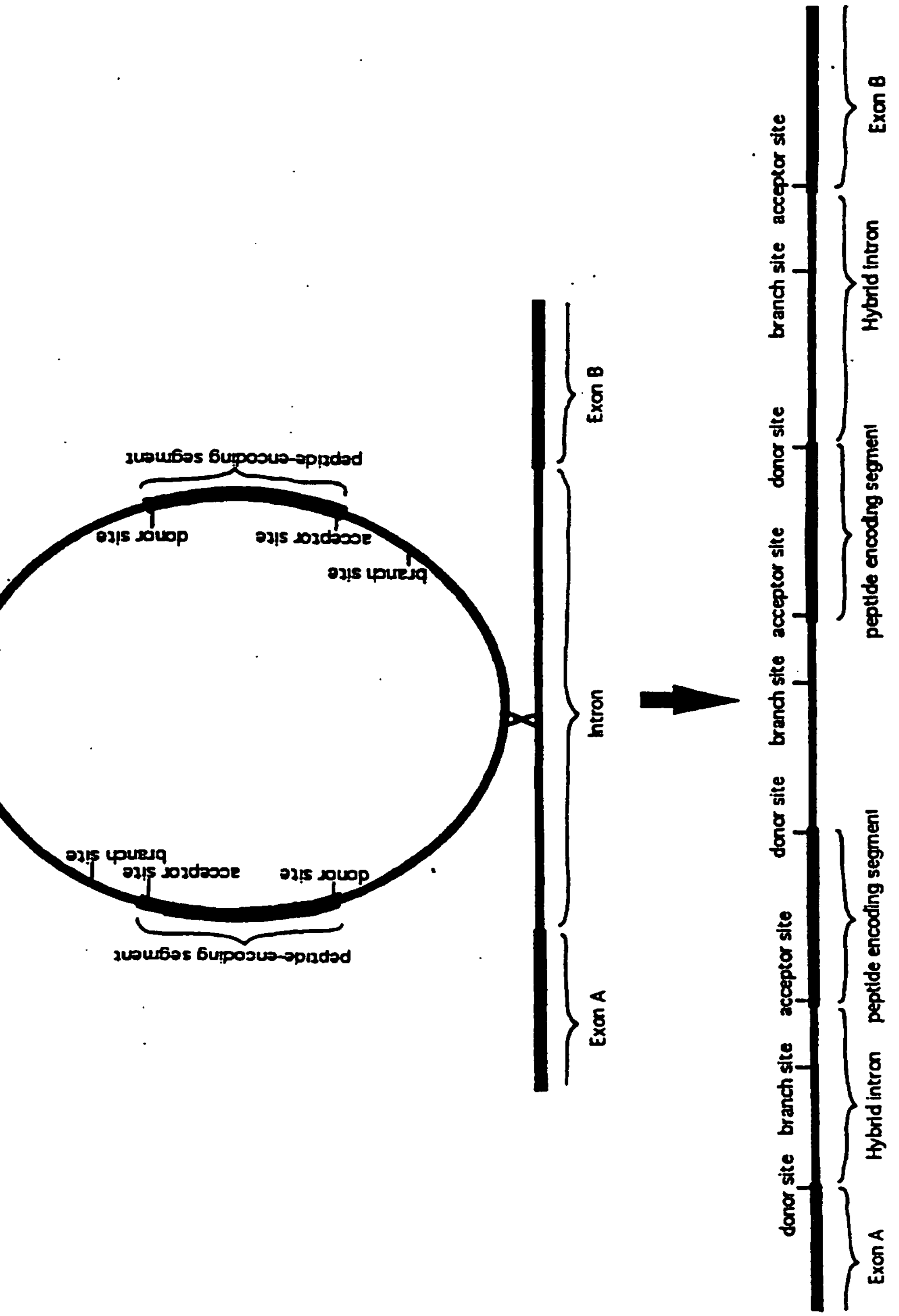
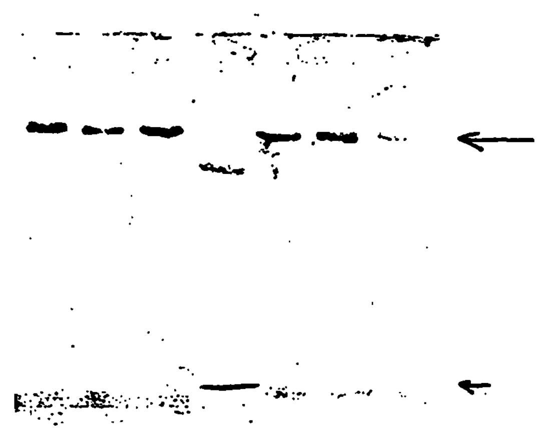


Figure 12: Structure of CD-DNA (embodiment 8) inserted into intron.

Figure 12: Western blot probed with monoclonal antibody 12CA5.



Lanes 1-3 and 5-7 contain total protein from six different *Chlamydomonas* transformants containing the CD-tagged RSP3 gene. Lane 4 (positive control) contains total protein from a yeast strain with a 15 kD epitope-tagged protein. The large arrow indicates the position of RSP3, and the small arrow indicates the position of the control protein.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US94/00254

A. CLASSIFICATION OF SUBJECT MATTER

IPC(5) :C12P 21/06,19/34; C12N 15/00,5/00,1/00.
US CL : 435 69.7,91,172.3,240.2 and 317.1.
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435 69.7,91,172.3,240.2 and 317.1.

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Dialog

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Gene, Volume 120, issued 1992, S. Sugano et al, "Use of an Epitope-Tagged cDNA Library to Isolate cDNAs Encoding Proteins with Nuclear Localization Potential", pages 227-233, see entire document.	1-3,5,8-25 and 33
Y	Gene, Volume 114, issued 1992, U. Pati, "Novel Vectors for Expression of cDNA Encoding Epitope-Tagged Proteins in Mammalian Cells", pages 285-288, see entire document.	1-3,5,8-25 and 33
Y	Nucleic Acids Research, Volume 20, No. 19, issued 1992, K. Luehrsen et al, "Insertion of Non-Intron Sequence Into Maize Introns Interferes with Splicing", pages 5181-5187, see page 5184, col. 1, parag. 2 to page 5185, col. 1, line 13.	1-3,5,8-25 and 33

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be part of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
04 APRIL 1994	APR 19 1994
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer Deborah Crouch, Ph.D. Jill Warden for Telephone No. (703) 308-0196

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US94/00254

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	EMBO Journal, Volume 3, No. 13, issued 1984, S. Munro et al, "Use of Peptide Tagging to Detect Proteins Expressed from Cloned Genes: Deletion Mapping Functional Domains of *Drosophila* hsp70", pages 3087-3093, see entire document.	1-3,5,8-25 and 33

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US94/00254

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

Please See Extra Sheet.

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-3,5,8-25 and 33

Remark on Protest
☐ The additional search fees were accompanied by the applicant's protest.
☐ No protest accompanied the payment of additional search fees.

BOX II. OBSERVATIONS WHERE UNITY OF INVENTION WAS LACKING
This ISA found multiple inventions as follows:

I. Claims 1-3,5,8-25 and 33, drawn to a method of tagging genes, a eukaryotic cell containing genes tagged by the method and a macromolecule tagged by the method, classified in 435, subclasses 91,172.3 and 240.2; Class 530, subclass 350 and Class 536, subclass 27.

II. Claims 1-4,10-23 and 33, drawn to a method for tagging genes and a prokaryotic cell microorganism containing genes tagged by the method, classified in Class 435, subclasses 91,172.3 and 252.3.

III. Claims 1-3,6,10-23 and 33, drawn to a method for tagging genes and a multicellular organism (animal) containing genes tagged by the method, classified in Class 435, subclass 91,172.3 and 240.2 and Class 800, subclass 2.

IV. Claims 1-3,6,10-23 and 33 are drawn to a method for tagging genes and a multicellular organism (plant) containing genes tagged by the method classified in Class 435, subclasses 91 and 172.3 and Class 800, subclass 200.

V. Claims 1-3,7,10-23 and 33, drawn to a method for tagging genes and a multicellular organism (animal) produced as a result of gene therapy and containing genes tagged by the method, class 435, subclasses 91 and 172.3; Class 514, subclass 44 and Class 800, subclass 2.

VI. Claims 1-3,7,10-23 and 33 drawn to a method for tagging genes and a multicellular organism (plant) produced as a result of gene therapy and containing genes tagged by the method, classified in Class 435, subclasses 91 and 172.3; Class 514, subclass 44 andClass 800, subclass 200.

VII. Claims 26 and 27,drawn to a subcellular extracellular structure, classified in Class 435, subclass 317.1.

VIII. Claims 28 and 29, drawn to a diagnostic test for a disease, classified in 435, subclasses 6 and 91.

IX. Claim 30, drawn to a therapeutic treatment for a disease, classified in Class 514, subclass 44.

X. Claims 31 and 32, drawn to an instrument, classified in Class 435, subclass 287.

XI. Claim 34, drawn to a DNA molecule, classified in Class 536, subclass 27.

The inventions of groups I-VI are to patentably distinct species because the methods required for the transformation of prokaryotic cells, eukarytoic cells, intact animals and intact plants are separate and independent. In addition the methods of germ cell transformation as in group I-IV are distinct and independent of the methods for gene therapy. In addition one is not need for the other. The inventions of groups VII,VIII ,IX,X and XI are drawn to mutually exclusive and independent products. There is no obvious relationship between a macromolecule, a diagnostic test for a disease, a therapeutic treatment for disease, an instrument and a DNA molecule. Thus the claims of these groups are drawn to distinct inventions which are not so linked as to form a singe inventive concept. PCT Rule 13.1 and 13.2 do not provide for multiple products and methods.

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: ______________________

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.